AUTOMATED MORPHOLOGICAL ANALYZER GENERATOR OF ANGIKA

SYNOPSIS

Submitted in partial fulfillment of the requirements

for the award of the degree of

Doctor of Philosophy

in the

Department of Computer Science & Engineering

Submitted to

YBN University

YBN UNIVERSITY

RANCHI

JHARKHAND – 834010, INDIA

Jun, 2021

Research Supervisor:
Dr. Manisha Kumari Deep
Professor, School of Computer Science & IT
YBN University, Ranchi

Research Scholar:
Alok Kumar
Department of Computer Science & Engineering
YBN University, Ranchi
Enrollment No.: YBNU2003277

Title of the Research: Automated Morphological Analyzer Generator for Angika

1. Introduction

Natural Language Processing (NLP) is an interdisciplinary field that focuses on enabling computational systems to analyze, understand, and generate human languages. One of the most fundamental components of NLP is morphological analysis, which involves identifying the internal structure of words by decomposing them into roots, affixes, and grammatical features such as tense, number, gender, person, case, aspect, mood, and honorificity.

Morphological analysis plays a crucial role in the processing of morphologically rich languages, where a single lexical root can generate a large number of surface word forms through inflectional and derivational processes. Most Indian languages belonging to the Indo-Aryan family exhibit complex morphological systems, making morphological analysis a prerequisite for effective computational processing.

Despite the linguistic richness of Indian languages, the development of computational resources has remained uneven. While high-resource languages such as Hindi have benefited from extensive NLP tools and annotated corpora, many regional and minority languages continue to remain underrepresented in computational linguistics.

Angika, an Eastern Indo-Aryan language spoken primarily in the states of Bihar, Jharkhand, parts of West Bengal, and the Terai region of Nepal, is one such low-resource language. Although Angika possesses a systematic and rule-governed morphological structure, it has received very limited attention in NLP research. Essential computational resources such as morphological analyzers, annotated corpora, and lexical databases are largely absent for the language.

A morphological analyzer functions as a core linguistic tool that supports several higher-level NLP applications, including part-of-speech tagging, syntactic parsing, machine translation, information retrieval, and speech processing. For low-resource languages like Angika, the development of a morphological analyzer is not only a technical necessity but also a linguistically significant step towards language documentation, preservation, and digital inclusion.

In this context, the present research aims to design and develop an Automated Morphological Analyzer Generator for Angika using a linguistically grounded and rule-based computational approach. The study seeks to bridge the gap between descriptive linguistic analysis and computational implementation, thereby contributing to both theoretical linguistics and applied NLP research.

2.Need of the Study

The need for the present study arises from several interrelated linguistic and computational gaps associated with the Angika language.

Firstly, Angika lacks essential NLP tools such as morphological analyzers, part-of-speech taggers, and annotated corpora. The absence of these foundational resources restricts the development of advanced language technologies and limits the inclusion of Angika in modern digital platforms.

Secondly, Angika is a morphologically rich language exhibiting complex inflectional and derivational processes across lexical categories such as nouns, verbs, adjectives, and postpositions. Without automated morphological analysis, this linguistic richness remains underutilized in computational applications.

Thirdly, existing NLP research on Indian languages has predominantly focused on high-resource languages, leading to the marginalization of regional and minority languages like Angika. Tools developed for Hindi cannot be directly adapted to Angika due to significant differences in morphological structure and usage.

Furthermore, data-driven and machine learning–based approaches are often unsuitable for low-resource languages due to the lack of large-scale annotated datasets. In such contexts, linguistically interpretable and rule-based computational models offer a more sustainable and reliable solution.

Finally, the development of computational tools for Angika contributes to the digital preservation and empowerment of the language, enabling its use in education, governance, and digital communication. Therefore, the present study is both academically relevant and socially significant.

3. Problem Statement

Despite its linguistic richness and structural regularity, Angika remains severely underrepresented in the domain of Natural Language Processing. One of the most fundamental challenges in developing computational tools for Angika is the absence of an automated morphological analysis system capable of handling its inflectional and derivational complexity.

Angika words undergo multiple morphological transformations based on grammatical categories such as number, gender, person, tense, aspect, mood, case, and honorificity. A single lexical root can generate a wide range of surface forms through the addition of suffixes, postpositions, and morphophonemic alternations. However, these morphological processes are currently described only at a descriptive linguistic level and have not been systematically formalized for computational processing.

The lack of a morphological analyzer for Angika leads to several interrelated problems. There is no automatic mechanism to identify roots, affixes, and grammatical features from surface word forms. As a result, higher-level NLP applications such as part-of-speech tagging, syntactic parsing, machine translation, and information retrieval cannot be reliably developed for the language. Moreover, data-driven approaches are impractical due to the absence of large annotated corpora.

The present research addresses this gap by proposing the design and development of an Automated Morphological Analyzer Generator for Angika, which systematically formalizes linguistic rules and implements them through a computational framework suitable for low-resource language settings.

4. Objectives of the Research

The primary objective of the present research is to design, develop, and evaluate an Automated Morphological Analyzer Generator for Angika that is linguistically interpretable, computationally efficient, and suitable for a low-resource language environment.

To achieve this overarching objective, the study is guided by the following specific objectives:

1. To conduct a systematic and detailed linguistic analysis of Angika morphology, covering major lexical categories such as nouns, verbs, adjectives, and postpositions.

2. To identify and classify inflectional, derivational, and morphophonemic processes operating in Angika, with special attention to grammatical categories such as number, gender, person, tense, aspect, mood, case, and honorificity.

3. To formally represent the identified morphological processes of Angika using explicit linguistic rules that can be computationally implemented.

4. To design an automated morphological analyzer generator capable of producing a functional morphological analyzer based on a defined lexicon and a set of linguistic rules, rather than constructing a static analyzer limited to a fixed dataset.

5. To develop a computational model using a rule-based and finite-state approach that ensures transparency, interpretability, and efficiency in morphological analysis.

6. To construct a dedicated Angika lexical database consisting of root forms and associated grammatical features required for morphological processing.

7. To address morphological ambiguity and surface-level variation through systematic rule ordering and controlled computational strategies.

8. To evaluate the performance of the proposed morphological analyzer using manually verified linguistic data, focusing on accuracy in root identification, morpheme segmentation, and feature assignment.

9. To contribute a reusable and extensible computational framework that can support further NLP research and application development for Angika and other linguistically similar low-resource Indian languages.

5. Research Questions

The present study seeks to address the following research questions:

1. What are the major morphological processes governing word formation in Angika across different lexical categories?

2. How can the inflectional, derivational, and morphophonemic patterns of Angika be systematically formalized for computational processing?

3. To what extent can a rule-based and finite-state approach effectively model the morphological complexity of Angika in the absence of large annotated corpora?

4. How can a morphological analyzer generator be designed to ensure flexibility, reusability, and linguistic transparency for a low-resource language like Angika?

5. How accurately can the proposed system analyze Angika surface word forms in terms of root identification, morpheme segmentation, and grammatical feature assignment?

6. What is the potential contribution of the developed morphological analyzer generator to higher-level NLP applications and digital language preservation for Angika?

6. Scope of the Study

The scope of the present research is confined to the linguistic and computational analysis of Angika morphology within a rule-based Natural Language Processing framework.

The study focuses on the following aspects:

- The analysis of major lexical categories in Angika, including nouns, verbs, adjectives, and postpositions, with respect to their morphological behavior.
- The examination of inflectional and derivational processes relevant to grammatical categories such as number, gender, person, tense, aspect, mood, case, and honorificity.
- The formalization of morphological rules and morphophonemic alternations required for computational analysis.
- The design and implementation of an automated morphological analyzer generator using a finite-state and rule-based approach.
- The development of a lexical database and rule engine specifically tailored for Angika.

The study does not aim to develop large-scale annotated corpora or employ data-driven machine learning models. Instead, it deliberately emphasizes linguistically interpretable and resource-efficient computational methods suitable for low-resource language contexts.

7. Research Methodology

The present research adopts a linguistically grounded, rule-based computational methodology for the design and development of an Automated Morphological Analyzer Generator for Angika. Considering the low-resource nature of the language and the absence of large-scale annotated corpora, the study deliberately avoids a purely data-driven or machine learning–based approach. Instead, it integrates descriptive linguistic analysis with formal computational modeling, ensuring transparency, interpretability, and sustainability.

The overall research methodology is divided into the following interrelated phases:

7.1 Linguistic Data Collection and Analysis

The first phase of the research involves the systematic collection and analysis of Angika linguistic data. The data is obtained from multiple sources, including native speaker intuitions, written texts, and existing descriptive linguistic studies on Angika.

Special emphasis is placed on collecting representative word forms across different grammatical categories. A detailed linguistic analysis is carried out to identify:

- Root forms and base morphemes
- Inflectional suffixes related to number, gender, person, tense, aspect, mood, and case
- Derivational morphemes resulting in category change or semantic modification
- Morphophonemic alternations occurring at morpheme boundaries
- Patterns of honorificity and agreement, particularly in verbal morphology

This phase provides the linguistic foundation necessary for computational formalization.

7.2 Morphological Classification and Feature Specification

In the second phase, Angika words are classified into major lexical categories such as nouns, verbs, adjectives, and postpositions. Each category is further analyzed to identify relevant morphological features associated with it.

For example:

- Nouns are analyzed with respect to number, gender, case, and postpositional attachment.

- Verbs are analyzed in terms of tense, aspect, mood, person, number, and honorificity.
- Derivational processes are examined to understand the formation of new lexical items from existing roots.

This classification enables the structured representation of linguistic information required for computational processing.

7.3 Rule Formalization

Based on the linguistic patterns identified, the next phase involves the formalization of morphological rules. These rules define how surface word forms are generated from underlying lexical representations.

The rules specify:

- Valid combinations of roots and affixes
- Constraints on morpheme ordering
- Conditions governing morphophonemic alternations
- Feature compatibility across morphemes

The formalized rules are designed to be explicit, modular, and reusable, making them suitable for computational implementation.

7.4 Justification for Rule-Based and Finite-State Approach

The choice of a rule-based and finite-state approach is motivated by both linguistic and practical considerations. Angika is a low-resource language for which large annotated corpora are unavailable, making data-driven and neural approaches impractical.

Finite-state methods are particularly well-suited for morphological analysis as they efficiently model concatenative and morphophonemic processes while maintaining computational efficiency. Moreover, rule-based finite-state systems offer linguistic transparency, allowing each analysis to be directly traced back to explicit linguistic rules.

This approach ensures that the developed system is interpretable, resource-efficient, and adaptable to future extensions, thereby making it appropriate for Angika and other similar low-resource languages.

7.5 Design of the Morphological Analyzer Generator

Unlike a static morphological analyzer, the proposed system is designed as a morphological analyzer generator. This generator produces a functional analyzer based on:

- A defined lexical database
- A set of formalized morphological rules
- Feature specifications associated with lexical items and affixes

This design allows linguistic knowledge to remain separate from processing logic, enabling easy modification, extension, and reuse of rules without restructuring the entire system.

9. Expected Outcomes of the Study

The present research is expected to produce the following concrete and academically significant outcomes:

- A linguistically well-defined morphological description of Angika, covering major lexical categories and their inflectional, derivational, and morphophonemic processes.
- A formally specified set of computationally implementable morphological rules that systematically capture the morphological structure of Angika.
- The development of an Automated Morphological Analyzer Generator for Angika, capable of generating a functional morphological analyzer based on a defined lexicon and rule set.
- A working morphological analyzer that can accurately identify roots, segment morphemes, and assign grammatical features such as tense, number, gender, person, case, aspect, mood, and honorificity.
- A dedicated Angika lexical database consisting of root forms and associated grammatical features required for morphological analysis.
- An evaluation of the proposed system demonstrating its effectiveness in handling Angika word forms with acceptable accuracy in a low-resource setting.

- A reusable and extensible rule-based computational framework that can be adapted for other linguistically similar low-resource Indian languages.

10. Significance and Contribution of the Study

The significance of the present research lies in its contribution to linguistic theory, computational linguistics, and low-resource language technology.

From a linguistic perspective, the study provides a systematic and formalized account of Angika morphology, thereby strengthening the descriptive and analytical documentation of the language. It bridges the gap between traditional linguistic description and formal computational representation.

From a computational perspective, the research contributes a transparent and interpretable rule-based model for morphological analysis, demonstrating the suitability of finite-state methods for morphologically rich and low-resource languages.

From the perspective of Natural Language Processing, the developed morphological analyzer generator serves as a foundational tool that can support higher-level NLP applications such as part-of-speech tagging, syntactic parsing, machine translation, and information retrieval for Angika.

From a socio-linguistic and cultural perspective, the study contributes to the digital preservation and technological empowerment of Angika by enabling its inclusion in modern language technologies, educational platforms, and digital communication.

Finally, the research contributes to the broader domain of low-resource language NLP by proposing a reusable and linguistically grounded framework that can be extended to other under-resourced Indian languages with similar morphological characteristics.

11. Tentative Chapter Scheme

The proposed research work is tentatively planned to be organized into the following chapters:

1. **Chapter 1: Introduction**

   This chapter will introduce the research problem, objectives, scope, and significance of the study. It will also provide an overview of Angika as a low-resource language and discuss the role of morphological analysis in Natural Language Processing.

2. **Chapter 2: Review of Literature**

   This chapter will review relevant linguistic and computational literature related to morphology, morphological analysis, finite-state methods, and NLP research on Indian and low-resource languages, with specific reference to Angika.

3. **Chapter 3: Morphological Structure of Angika**

   This chapter will present a detailed linguistic analysis of Angika morphology, covering nouns, verbs, adjectives, postpositions, inflectional and derivational processes, and morphophonemic alternations.

4. **Chapter 4: Formalization of Morphological Rules**

   This chapter will focus on the formal representation of Angika morphological processes, including rule specification, feature structures, and constraints required for computational modeling.

5. **Chapter 5: Design of the Morphological Analyzer Generator**

   This chapter will describe the conceptual design and architecture of the morphological analyzer generator, explaining its modular structure and functional components.

6. **Chapter 6: Implementation and Evaluation**

   This chapter will discuss the computational implementation of the proposed system, evaluation methodology, test data, experimental results, and analysis of system performance.

7. **Chapter 7: Conclusion and Future Directions**

   This chapter will summarize the findings of the research, highlight the major contributions of the study, discuss limitations, and suggest possible directions for future research.

12. Timeline

The proposed research work is planned to be completed within the prescribed duration of the PhD programme as follows:

Period	**Work Description**
First 6 months	Review of literature and refinement of research framework
Next 6 months	Linguistic data collection and morphological analysis of Angika
Next 12 months	Rule formalization and design of the morphological analyzer generator
Next 6 months	Computational implementation and system development
Next 6 months	Evaluation, result analysis, and system refinement
Final 6 months	Thesis writing, revision, and submission